CES/SEM.46/WP.1
17 May 2001

ENGLISH ONLY

STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

COMMISSION OF THE EUROPEAN COMMUNITIES (EUROSTAT)

CONFERENCE OF EUROPEAN STATISTICIANS

Joint ECE/EUROSTAT Seminar on Business Registers
(Geneva, 28-29 June 2001)

CLAMOUR: IMPROVING THE QUALITY OF EXISTING AND FUTURE CLASSIFICATION SYSTEMS

Working paper prepared by Office for National Statistics of United-Kingdom*

Summary

The CLAMOUR project is partly financed under the European Union Fifth Framework Research and Technological Development Program. Its aims are to improve the quality of classification coding and to enhance the development of future classification systems by providing a foundational basis on which they can be designed. Methods for businesses to transmit information to governments in a structured, user-friendly way are being developed. The work of the project is underpinned by a thorough examination of users needs, which target the areas in need of improvement. This paper provides an overview of the project and its emerging findings, particularly concentrating on the users needs.

Keywords: Classifications, NACE, quality, user needs.

1. CLAMOUR (Classifications Modelling and Utilities Research), is a project partly financed under the European Union Fifth Framework Research and Technological Development Program.[1] The project aims to find a better way to classify when collecting and disseminating industrial information & statistics and to ensure that, as European countries work closer together, the collection and management of data is consistent.

---

* Prepared by Mr. Steve Vale, Office for National Statistics, Newport, UK.

2. The project, which focuses on industrial classification systems and on the EU system NACE in particular, started in January 2000 and ends in December 2001. The NSIs of France, Denmark, Finland, Netherlands and the UK, together with a French business and independent consultant form the consortium. Its main themes are:

- Users' Needs
  This has established areas of greatest concern for users and ensures the project delivers practical and valuable results
- Foundations
  A model has been developed to describe the structure and activities of business, providing a fundamental basis for defining statistical units and classifications, improving consistency in classifications and enabling flexibility and the construction of alternative classifications
- Data Providers' Questionnaire
  A questionnaire, written in Blaise, has been developed which implements the foundational model and allows businesses to communicate classification data to government by electronic means
- Linguistics
  Tools are being developed to improve the quality of classification coding through natural language processing.

This paper outlines the major objectives of the project and the work done to date.

User Needs

3. Statistical information about businesses is collected and classified so that accurate pictures of national economies can be produced. Government and research organisations also use this sort of data to inform policy decisions.

4. The user needs part of the CLAMOUR project aims to establish how information is used and what form of classification provides the best value for users[2]. Various National Statistical Institutions (NSIs) have interviewed a sample of users of government information. Questions ask how information is currently used and what changes in use may occur in the future.

Organisation

5. The user needs work package is co-ordinated by the UK Office for National Statistics (ONS). There has been little other work in the field of assessing how statistical classifications meet user needs, so considerable work was needed at the planning stage to give the survey a sound methodological basis. The main methodological input to this work was the report of an internal Eurostat project to find ways to assess the quality of statistical norms[3]. Certain decisions were necessarily taken on the basis of very limited information and experience in this field.

6. The survey was carried out using a questionnaire[4] designed and tested by ONS staff, in conjunction with the project partners in other countries. It was decided to concentrate on three main

areas of investigation:

- Current use of business and economic information and business classification schemes
- Future needs of business and economic information and business classification schemes
- Administration of classifications

7. An optional supplementary set of questions[5] was produced to gather additional information on preferences with respect to statistical units. It was felt that the issue is so complex that, instead of asking respondents their opinion on various definitions, case scenarios were constructed. The answers to these will be analysed in the foundations part of CLAMOUR, and will be used to define a number of statistical units that comply with user wishes.

8. Four categories of users were identified; private businesses, central government, local authorities and researchers. Users in each category were approached and invited to submit their responses.

9. There were two rounds of data collection, which are described below. A report describing the first round of data collection (project deliverable 14) was published on the Internet in February 2001[6], whereas the full results of the two rounds will be included in deliverable 15, to be published in June 2001.

Round One

10. The first round of data collection, involving the four project partner countries, Denmark, Finland, The Netherlands and the UK, started in the summer of 2000. The data collection phase was completed, and analysis of the results started in early 2001. The number of responses by group of users and country is shown in the following table.

	Denmark	**Finland**	**Netherlands**	**UK**	**Total**
Private Businesses	14	25	12	14	**65**
Central Government	14	11	20	25	**70**
Local Authorities	6	1	5	8	**20**
Research	6	5	8	3	**22**
Total	**40**	**42**	**45**	**50**	**177**

Round Two

11. Twenty European countries were invited to take part in the second round of data collection using a simplified version of the original questionnaire, with a target response of eight completed forms

per country. The data collection took place between December 2000 and February 2001. Responses were received from eleven countries. The same questionnaire was also sent to the statistical part of various international organisations. Responses were received from the OECD and the UN/ECE.

Emerging Findings and Links with Other Work Packages

12. Detailed results of the user needs survey will be included in a future report (Project Deliverable 15 - "Final Report Following the User Survey in other European Countries" - due in summer 2001). Initial analysis allows us to identify a few broad themes where the responses are giving clear messages:

- Most users want a higher level of detail in future classifications, particularly for service activities;
- Many users would like future classifications to have more of a product focus, concentrating more on outputs rather than purely on activities;
- Better coverage of new and developing activities is needed. A large number of users specifically identified information, communication and technology (ICT) activities as an area where existing classifications need to be improved. More guidance on classifying new activities would be useful, perhaps including an interactive case-law database;
- Users want "better" or more flexible ways of aggregating codes within classifications, to allow better analyses of specific economic phenomena. Multi-dimensional classification systems may help to meet the needs of more users;
- More flexible unit structures are needed to better meet the needs of all users. This implies that data should be collected at the lowest feasible level within a business, and then aggregated in various ways to meet different needs;
- A significant minority of users place a strong emphasis on the legal unit view. They want statistical units to follow legal structures. This view varies between countries, and is strongest in the other European countries included in the second wave of the survey, the UK, and to a slightly lesser extent in Denmark;
- Many users make a strong distinction between market and non-market activities. Some want to identify the non-market activities within a business, whereas others see them as irrelevant. This implies that data collection systems need to make a clear distinction between these types of activity. Holding companies are a specific case, which require careful treatment and clearer guidelines.
- The needs of users seem to be broadly consistent across participating countries
- These emerging findings are being fed into CLAMOUR Work Packages six, seven, eight, ten and eleven, where they will be used to develop models and tools for collecting business information.

Initial conclusions regarding the data collection exercise

13. The questionnaire has proved to be a useful tool to collect information on the needs of users regarding business classifications, however, there is scope to refine and improve it in the light of experience if it is to be used again.

14. The English language version of the questionnaire was not always as clear as it might have been, even causing difficulties for some of the UK respondents. A re-write to simplify and clarify the language would have helped improve understanding, and made translation much easier.

15. The use of different software packages and different data formats in the participating countries meant that data exchange was not always as efficient as it should have been. In future these issues should be addressed and resolved at the outset.

16. The development of a multi-lingual web-based questionnaire tool, possibly using Blaise software, would have facilitated data collection. This was outside the scope of this Work Package, but should be considered for any future exercises of this type.

17. The results of this exercise provide a very useful input into future revisions of NACE, and other types of business classifications. The approach used could be transferred to other classification systems, and possibly to other statistical norms.

Foundations

18. The major aim of the foundational sub-project is to develop a model to describe the structure and activities of a business. One version of the model will be comprehensive enough to satisfy all anticipated users needs. A simplified model is then derived from this, which can be built using the information businesses, can realistically expect to provide us. The model is described in greater detail in the paper to be presented by Jacobiene van der Hoeven7, so it shall only be touched on briefly here.

Uses of the Foundational model

19. The potential benefits of the model are various. Firstly it will enable the development of multi-purpose business registers; so that structured information about businesses can be used in various ways. One application of such flexibility is the ability to code businesses in many ways, producing statistics about crosscutting industry sectors such as tourism or the environment.

20. The model should enable the construction of statistical units of various types, with the use of algorithms ensuring a greater harmonisation within and between countries. Algorithms can then be used to classify these units to classification codes, again improving the quality of coding.

21. The model may also be used as a framework for integrating administrative data in statistical data collection systems, possibly in relation with a single business register. It also provides a language in which national statistical systems can be compared and international harmonisation improved. It may be used in the definition and measurement of the quality of statistical units and classifications. And, perhaps most important, it fosters flexibility of statistical systems at a time when the existing statistical framework struggles with the demand for data on the changing economy.

Building the model

22. In the model a number of entities are defined, and for each entity a number of attributes. The smallest entity of the model is called the building block, from which activities and structures can be built. A business typically consists of a large number of building blocks, which are derived by splitting a business into parts, each of which relates only to a single location, legal unit, accounting system, owner, etc. Statistical units can then be derived by clustering building blocks by the application of an aggregation algorithm, which is essentially a function of the attributes of the building blocks.

23. Care is taken that the model can generate statistical units and classifications in accordance with the needs of the users of statistical data, and that applying the model in practice is feasible. The user needs survey has provided input to the model. Additionally, the Data Providers Questionnaire described below is being developed for the collection of data needed for the application of the algorithms for forming statistical units and classifying them. The data will be used to test the algorithms.

Data Providers' Questionnaire (DPQ)

24. The objective of the DPQ is to develop an electronic questionnaire which is able to gather the information required by the Foundational model, in a structured format, and in a way which can be used easily by businesses to communicate with governments. The questionnaire has been developed using Blaise software in a sub-project led by CBS in the Netherlands. A first version of the DPQ was tested in 3 partner countries; Finland, UK and the Netherlands, and those countries will test a second version in June. A demonstration system will then be produced by the end of 2001.

Version 1 of the DPQ

25. Blaise was used to implement the questionnaire as it is suitable for conducting surveys and allows for multi-lingual questionnaires to be developed. As one of the objectives in the project is to ensure greater consistency in classifications, this was an important consideration. The questionnaire is structured in a hierarchy, with the aim that when sufficient information has been gathered, there is no need to continue with further detail. Closed questions are used where possible, to structure the information gathered, although the use of free-text responses is necessary for certain questions. Questions cover the information needed for the model; inputs, outputs, processes, materials, customers, etc.

26. The main quality criteria for the DPQ is that national 5-digit derivations of NACE could be obtained from the information gathered. The questionnaire also has to be practical, enabling it to be used by all kinds of businesses.

27. For the first version, one industry sector was chosen; the wood industry. This means that code lists, from which answers are selected only had to be constructed for that particular sector. It was also considered useful for the first version to use a relatively simple industry, where the range of activities was not too complex.

Testing the DPQ

28. Each of the 3 countries approached a variety of businesses in the wood industry to test the DPQ on a voluntary basis and provide feedback on its ease of use. A variety of interviewing methods was used, which allows for comparison between them as to their relative suitability. The Netherlands conducted personal interviews with the use of a laptop. The UK posted and emailed discs to businesses. Finland conducted interviews by telephone.

29. Testing highlighted several areas for improvements, and provided useful information on how later versions could be improved. It was considered beneficial to include known information about the business in advance, to reduce the burden on them. The questionnaire could be used to check the accuracy of this data. It was considered important to structure look-up tables so that the respondent is directed to correct level needed for the 5-digit NACE classifications: there was a danger that either too much or too little information is gathered.

30. Testing has proved that the questionnaire is a viable means of obtaining data needed to classify businesses, although there are improvements to be made. Generally businesses found the questionnaire not too difficult to complete, although it is considered there may be problems when tested on more complex businesses or industry sectors. Where a business has a wide range of products, or complicated production processes, it becomes increasingly difficult to provide detailed information, such as turnover for each product.

31. Some work will need to be done on the usability of the questionnaire; to improve its look and feel and to provide users with information on its structure while they are completing it.

32. The use of Blaise as the means to implement the questionnaire has raised some discussion. It appears to work well where respondents are visited personally, or talked through the questionnaire by telephone. It is not considered very user-friendly to expect users to deal with computer discs or email files. In the light of this, the project is considering the possibility of using the web to complete the questionnaire, although this will raise its own set of problems to overcome.

Linguistics

33. The final sub-project is concerned with the development of linguistic tools which will enable the full meaning of business descriptions to be captured using natural language processing. While the foundational model and DPQ can be considered to use a closed system, the linguistic research tackles the problem of classification from an alternative, open viewpoint. Although there is some potential to integrate the two approaches, it is more reasonable to view them as complimentary, likely to be used in different circumstances.

34. This sub-project is being led by INSEE, with the technical development being done by LexiQuest, a French linguistic company. The linguistic research is building on LexiQuest's existing system for processing natural language queries, which can be used as an interactive web application or

in batch mode.

35. The main objective is to improve the quality of classification matches to produce a reliable system. Three issues are being researched to address this, together with research into tools which will assist matching between classifications and develop semi-automatic correlations.

The three main areas of research

- Context-based semantic disambiguation
- This aims to extract the true meaning of a word based on its context. Research will identify ambiguous entities and specify procedures by which they can be dealt with, using information on their domains.
- Compound term processing
- This workpackage will aim to produce rules for identifying compound terms based on their linguistic properties, deciding whether a compound term should be indexed for future retrieval, and whether a particular component should be promoted.
- Highly structured descriptions
- This work will develop procedures by which complicated descriptions may be simplified, identifying the elementary components, and annotated with information on their relevance.

Languages and progress

36. Inconsistencies between languages are one of the main reasons behind the project, so it is important to maintain consistency in the tools as they apply to different languages. Four languages (English, French, Spanish and German) are researched in depth, together with Italian and Dutch to a lesser degree, and results will be compared to ensure comparability between results in the different languages. Work on the main languages in all of the research areas is being undertaken in parallel, to be followed with the others.

Issues

37. A key success factor in the linguistic approach is the quality of the dictionaries used to underpin the system. It is particularly important where technical language is used, such as classification data, in which meanings of words differ from their popular meanings. Research to date has concentrated on general language usage, but will be focussing on the specific application for classifications.

38. The project will produce reports and algorithms demonstrating how a linguistic system may be introduced. However, it is anticipated that a substantial amount of work is needed for any NSI to do this, and that much of this work is involved in ensuring the dictionaries to be used are of sufficient quality. INSEE intend to implement the results for classification coding in France.

Conclusions

39. The project is aimed at delivering real benefits to the future maintenance and development of classification systems. The needs of users of classifications and the statistics on which they are built have been at the centre of our work. Work is generally progressing according to plan, and independent assessors have been reassured that it is of good quality.

40. The tools to be produced by the end of the project will be implemented to some degree by the project partners. However, to fully implement the findings a co-ordinated effort across Europe will be required. It is hoped that the revision to NACE in 2007 will offer the opportunity to do this.

NOTES

[1] For further information on CLAMOUR, visit the project website: http://www.statistics.gov.uk/nsbase/methods_quality/clamour.asp

[2] For more information see the "Users' Needs" part of the CLAMOUR web-site - http://www.statistics.gov.uk/methods_quality/ClamourUsersNeeds.asp

[3] Eurostat sub-group - Quality of Norms, Final Report, July 1998.

[4] See http://www.statistics.gov.uk/methods_quality/downloads/questionnaire_letter.pdf

[5] See http://www.statistics.gov.uk/methods_quality/downloads/questionnaire_cases.pdf

[6] See http://www.statistics.gov.uk/methods_quality/downloads/Deliverable14.pdf

[7] *Modelling the Organisation and Economic Activities of Businesses*, Struijs, van der Hoeven and Kroese.